NEWS AND FINANCIAL MARKET

BY

JING ZOU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Professor ChengXiang Zhai

ABSTRACT

News can influent the market. It has been proven that using text mining techniques, financial news can be used to predict market trend and volatility. In this thesis, we study some existing prediction algorithms that are based on Naive Bayes Classifier and Adjusted Document Frequency-Inverse Document Frequency(ADFIDF) weighting. Although occurrence of features selected by ADFIDF weighting can usually represent volatility bursts in financial market, it has been unclear whether it is also effective for market trend or trading volume. We conduct experiments that test the effectiveness of ADFIDF feature selection algorithm in finding correlation between news, market volume and trend. Our experiement result shows that a thin positive correlation exists between ADFIDF features occurrence and market volume. However, features occurrence and market trend are not directly correlated. We also propose a novel algorithm of finding correlation between news and financial market. The proposed algorithm is based on topic model and adjust TF-IDF weighting. It allows us to identify a few factors that could influence the performance of a prediction algorithm, such as number of topics of a model and adjustment. of IDF value. Our experiment results also show that grouping stocks together does not necessarily improve the performance of a prediction algorithm.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Stock, established in 16th century, is ownership in part of a company. Usually, price of one company's stock is based on this company's financial performance. An investors who purchases stocks of a company believes that the company will do well and bring him or her fortune. Sometimes their wishes are fulfilled with growth of the company, while sometimes the stock certificate is worth no more than just a piece of paper. To hedge risks and maximize profit, many investors buy and sell different kinds of stocks frequently and only hold them for a short time. Nowadays, we call these investors "traders"

Since late 20th century, with the rapidly developed networking technology, Algorithmic Trading, a electronic platform based trading system, has attracted more and more people's attention. "In 2006 at the London Stock Exchange, over 40% of all orders were entered by algorithm traders" (Timmons, 2006) [1]. High Frequency Trading(HFT), a type of algorithmic trading that is characterized by short portfolio holding periods, accounts for 73% of all US equity trading volume(Haldane, 2010)[2]. In addition, financial products including future, option, equity and bond are also accessible to algorithmic traders. Traders now combines mathematical, statistical and computational techniques with business knowledge, to build automated algorithms to trade through electronic platform. The principle of these algorithms is trying to learn the market behavior, predict its trend and take action in the way that probabilistically should generate profit. In addition to market data histories, traders also need a variety of latest news information including financial, technology and sometimes even political information, to make accurate judgment about seemly unpredictable market. Dow Jones could fell dramatically due to financial panic caused by rise of unemployment rate. Then an excellent jobs report of United States could bring a rebound to the market[3](Hook & Lauricella, 2012). To capture important information that might influent the market, traders usually have an Bloomberg Terminal, a product that provides all kinds of latest news and data, on their desk while "baby-siting" their trading program. They might adjust their algorithm if there is some breaking news showing up that might cause abnormal trend or volatility

in the market. As speed and efficiency become more and more important in trading, it is interesting to question whether text mining and categorization techniques can capture key information in news? Further, using only news, is it possible to automatically predict market trend and volatility?

Previous studies have shown that news information indeed has impact on stock market. Gross-Klussmann and Hautsch(2011) examined high-frequency market reactions to an intraday stock-specific news flow[4]. They used linguistic pattern recognition and company-specific news to build a module that have "distinct responses in returns, volatility, trading volumes and bid-ask spreads due to news arrivals"(Gross-Klussmann, Hautschp, 2011). Researchers at The Chinese University of Hong Kong(2010) claimed that their "experiment showed a .4252 correlation between actual volatility and feature phrases' occurrence". The algorithm they proposed is using adjust document frequency inverse document frequency(ADFIDF) to select feature words[5]. Market volatility represents the instability of a market or stock. High volatility indicates that a lot attention is attracted. It is reasonable to assume that if a market has high volatility, then the volume should also be high. In their paper, they only test the co-occurrence between features and volatility burst while we hypothesize that features' occurrence should also be correlated with other market information. In this thesis, we conduct experiments and exam the effectiveness of ADFIDF feature selection in finding correlation between news, market volume and trend. Our experiement result shows that ADFIDF features' occurrence is correlated with market volume while not with market trend. The result support our hypothesis. Volume of the market is easier to predict. Whenever a stock is effected by a related news, the volatility will be increased because people might see potential profit or risk from the news information. Either interpolation will cause more trading activities, which increase the market volatility and trading volume. The direction of market in such situation can be therefore more random and unpredictable relatively.

Although there are many existing works about building prediction model based on news informations, using topic model to find correlation between news and market information is still a new approach. In this thesis, we propose an algorithm of finding correlation between news and financial market. The proposed algorithm is based on topic model and adjust TF-IDF weighting. It allows us to identify a few factors that could influent the performance, such as number of topics of a model and IDF adjustment. Our experiment result also shows that grouping stocks together does not improve the performance of prediction.

The remainder of this thesis is organized as follows. Chapter 2 gives account of related work and general challenges. The approaches to the problem and methodology are described in Chapter 3 folloewed by experimental study and results in Chapter 4. Finally, Chapter 5 will conclude my findings and discuss possible future improvement of research.

CHAPTER 2

RELATED WORK

In the business world, understanding sentiments and trends from news can have a great impact on business decision making. For example, one can get a glimpse how well the economy is doing by focusing on keywords such as GDP, unemployment, inflation, etc. When the economy seems to be heading upwards, corporations would start hiring more; retailing companies would stock up their inventory, or ramp up their manufacturing facilities; investors would pull their funds out of safe investments such as bonds and blue-chip stocks, and start investing in riskier assets such as small-cap stocks. Similarly, when sentiment turns and a downturn seems more imminent in the economy, all those trends mentioned above would reverse. Dow Jones could fell dramatically due to financial panic caused by rise of unemployment rate. Then an excellent jobs report of United States could bring a rebound to the market[3](Hook & Lauricella, 2012).

Many research have been done to find the signals in news. Market will react differently under influence of good and bad news. Koppel and Shtrimberg(2004) built a model based on lexical features that can "distinguish good news from bad news with accuracy of about 70%". Hafez(2009) claimed that "stocks tend to be underperforming prior to a negative news announcement" and "securities tend to outperform slightly prior to the release of a positive news event"[15]. He also found that "when bad news is released, the price impact is immediate" (Hafez, 2009). Not only financial related news are important to traders, news from many other area also have impact on the market. Anderson-Weir C. H.(2010) claimed that "the Unexpected Green Rankings had a significant effect on abnormal returns". One day after the 9/11 attack, the European and Latin American stock markets fell sharply(Fuerbringer & Norris). Because there is such a wide range of news that is able to influent the market, at this point, we cannot eliminate any kind of news when building prediction model.

Term frequency-inverse document frequency(TF-IDF) weighting is a common method of calculating how important a word is to a document in a collection. Both term frequency and inverse document frequency carry important information of a word in a documnet.

They can be used to select feature words from a financial news document. Then using regression algorithms, we can model the relationship between score of feature words and market information. Researchers at The Chinese University of Hong Kongy(2010) proposed a feature selection method base on adjust TF-IDF weighting, ADFIDD, to predict market volatility[5]. They claimed that there is a 0.4252 correlation between features occurrence and market volatility bursts(Pan, Cheng, Wu, Yu & Ke, 2010). Cheng, from Chienkuo Technology University(2010), used adjusted TF-IDF weighting and phrase-document matrix to predict trend a specific stock using Chinese news' articles[8]. He claimed that their prediction algorithm have nearly 100% accuracy when predicting one stock after 15, 30, 45 and 60 minutes after a news is released conference(Cheng, 2010). Although their algorithm has great performance, it haven't been examined using English news. Their research also raise a question: "Does the market only respond to news that are published with in 60 minutes?" To answer this question, in our research, we will use the difference between open and close price as market trend for each stock and predict market's daily trend and volume instead of hourly or shorter.

In 2010, researchers used Bayes algorithm to predict FTSEIOO, a share index of the stocks of the 100 companies listed on the London Stock Exchange with the highest market capitalization(Shihavuddin, Ambia, Ardin, Hossain, & Anwar, 2010)[9]. Naive Bayes Classifier is an algorithm that uses observations to calculate the likelihood of parameters. Experts of trading market construct a set of keywords which they think are important for moving markets. Words are grouped and assigned weights. Using Bayes classifier, they forecast in which class the analyzed message should be assigned to. Their algorithm had reasonable performance. The advantage of this method is that the implementation and logic are simple. However, this algorithm requires a large set of feature words with weights that represent their business values. The set is hard to get without financial experts and can not be standardized.

We can use different textual representations of news articles. Schumaker and Chen(2009) examines a predictive machine learning approach for financial news articles analysis using different textual representations, Bag of Words, Noun Phrases, and Named Entities[11]. Bag of Words is simply the standard news article with their stop words removed. Building upon the Bag of Words, Noun Phrasing is accomplished through the use of a syntax where parts of speech are identified through the aid of a lexicon and aggregated using syntactic rules on the surrounding parts of speech, forming noun phrases. Using semantic lexical hierarchy, named entities are obtained after classifying Noun Phrases into a person, organization, or location. Among all three representations, noun phrase has the

best directional accuracy at 50.7%(Schumaker and Chen, 2009). Named Entities have the best return, however it has the lowest directional accuracy(Schumaker and Chen, 2009). Names entities text representation might not be stable because names of important people and organizations are constantly changing.

CHAPTER 3

METHODOLOGY

3.1 Definitions

Market Trend A stock's trend of a trading day is the close stock price minus open price of that stock of the day. Market trend of $NASDAQ$ on a trading day is the close index minus the open index of the day.

Trading Volume Trading volume of a stock is the total quantity of contracts bought and sold during a trading day. Trading volume of $NASDAQ$ is sum of leading stocks' volume during a day.

Notations

1. d: a parsed news article.
2. $T = \{\mathrm{D}_1, D_2, ..., D_I\}$: collection of documents from day 1 to day I.
3. $D_i = \{d_1, d_2, ..., d_I\}$: document collection of day i.
4. $|D|$ = total number of documents in a collection.
5. $|w \in D|$ = number of documents that contain word w.
6. $M = \{m_1, m_2, ..., m_n\}$: market trend from day 1 to date n where m_i = close - open Dow Jones points of day i.
7. $S = [s_1; s_2; ...; sI]$: is a sequence of stock prices in the time interval I.
8. $mb_i = 1$ if close price of a stock or Dow Jones point minus open price or Down Jones point is greater than 0, -1 otherwise
9. $M_b = \{mb_1, mb_2, ..., mb_n\}$: binary market trend from day 1 to day n.

10. θ_{ij}: the topic distribution for document j on day i.

11. ϕ_k: is the word distribution for topic k.

12. $C(w|d)$: count of word w in document d.

3.2 ADFIDF

ADFIDF(Adjust Document Frequency Inverse Document Frequency) is a feature selection method proposed by researchers from The Chinese University of Hong Kong. ADFIDF value of stock k during interval I of a word w is defined as $ADFIDF(w|t) = \frac{|w \in D|}{|D|} \times log\frac{|D|}{|w \in D|}$ for all documents of time interval t. The ADFIDF measure does not account for term frequency because they discover that feature words have high coverage and low redundancy. In fact, Cheng(2010), also mentioned in his paper about market trend predicting that inverse document frequency is more important than term frequency. Using the co-occurrence rate to detect volatility bursts of a stock, they found a .4252 correlation between ADFIDF features occurrence and volatility bursts.

The volatility was defined as "standard deviation of the continuously compounded returns of a stock within a specific time horizon and is used to measure how widely prices are dispersed from the average"(Pan, Cheng, Wu, Yu & Ke, 2010). Bursty time intervals of the feature word f is a set of time that f appears, defined as represented by TB_f. The co-occurrence of feature and volatility was defined as $E(S; f) = \frac{V(S,TB_f)}{|TB_f|} / \frac{V(S,I)}{I}$, where $V(S; I)$ is the sum of the bursty volatility values regarding stock S in the time interval I.

In our research, we will be trying to find correlation between features, market trend and trading volume. So we definite the co-occurrence value as

$$E_{volume}(s; f) = \frac{Volume(s, TB_f)}{|TB_f|} / \frac{Volume(s; I)}{I}$$

$$E_{trend}(s; f) = \frac{Trend(s, TB_f)}{|TB_f|} / \frac{Trend(s; I)}{I}$$

where $Volume(s; I)$ is the sum of trading volume values and $Trend(s; I)$ is the trend sum regarding stock s in the time interval I. Then rank features by Algorithm1. F is collection of all words. γ is decay factor. To get correlation between trading volume, market trend and texture information. We define volume and trend estimate indices of documents of day D_i on stock S as, V_{est} and T_{est}, as

Algorithm 1 FeaturesRank (Pan, Cheng, Wu, Yu & Ke, 2010)

0: $\varepsilon \leftarrow \emptyset$
0: **while** $F \neq \emptyset$ **do**
0: f: $\forall$ f_j, $ADFIDF(f_j) \leq ADFIDF(f)$;
0: Remove f from F
0: $\varepsilon \leftarrow pair(f; E_{volume}(S; f))$;
0: **for all** $f_j \in F$ **do**
0: B = $TB_{f_j} \cap TB_f$
0: **if** $B \neq \emptyset$ **then**
0: $Volume(S; t) \leftarrow Volume(S; t) * \gamma$
0: update $E_{volume}(S; f_j)$
0: **end if**
0: **end for**
0: **end while**=0

$$V_{est} = \sum_{f_i} E_{volume}(S, f) * ADFIDF(f|i)/|f_i|$$

$$T_{est} = \sum_{f_i} E_{trend}(S, f) * ADFIDF(f|i)/|f_i|$$

3.3 Topic Model

Although there has been research done on finding a correlation between news and financial market, using topic model is still a now approach. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Each topic is a distributions over a set of words and each article has a distribution over a set of topics. For example, Topic("computer") maybe represented as {"hardware" 0.05, "CPU" 0.02, "software" 0.02, ...} and the article "Apple Releases New MacBook Air" has a topic distribution {"computer" 0.08, "technology" 0,05, "travel" 0.001, ...}. There are many algorithms of building a topic model. Latent Dirichlet Allocation(LDA) is probably the most commonly used algorithm. Gibbs sampling is a randomized algorithm used to approximate the joint distribution, in this case, the distribution of words in each topic. A major finding of this paper is that the more topics are built, the more stable and accurate the prediction will be. We explore the approaches of using topic model to find correlation between news information and stock market. After building the topic model, each article

has a distribution over topics. Market trend and volume are known. Assign each topic a score by simply summing up the product of probability of topics and market trend of the day for all documents. Normalize topic store. Then run a test article using topic model to get a distribution over topics and calculate score of this article. If this score is consistently correlated to market trend, then pattern exists between this feature and market information.

To see if topic model can help predicting market trend, first build a basic LDA topic model with Gibbs sampling using news articles. The model contains a list of topics $\{t_i, t_2, ..., t_k\}$. Calculate a score for each word by summing up and normalize the products of TF-IDF value and market trend of that day for all documents. After building the topic model, each article has a distribution over topics. Market trend and volume are known. Assign each topic a score by simply summing up the product of probability of topics and market trend of the day for all documents. Normalize topic store. Then Run a test article using topic model to get a distribution over topics. Calculate score of this article by summing up the product of probability of topics and topics' scores.

$$W(w_s) = \sum_i \sum_j TF - IDF(w_s|dij) * m_i/|D_i|$$

$$= \sum_i \sum_j \frac{C(w_s|d_{ij}) + 1}{|d_{ij}| + 1} * \log(\frac{|D|}{|d_{ij} \in D : w_s \in D_{ij}|}) * m_i/|D_i|$$

To classify an article d_{ij}, first calculate the score of the article by summing up the product of it's distribution of topics and topics' scores as: $S(d_{ij}) = \sum_t P(t|\theta_i j) * T(t)$. Consider two types of predicted market trends, binary and quantitative. Binary market trend is "+" if $S(d_{ij}) > 0$ and "-" if $S(d_{ij}) < 0$. Quantitative trend is simply the value of $S(d_{ij})$.

$$\text{Binary: } T_1(t_k) = \sum P(w|\phi_k) * W(w)$$

$$\text{Quantitative: } T_2(t_k) = \sum_{d \in D} P(t_k|\theta_{ij}) * m_i$$

To exam the effectiveness of models with different number topics, we build three models with 5, 10 and 15 topics.

CHAPTER 4

EXPERIMENT AND FINDINGS

4.1 Data & Tools

Market Data The primary data source of market information is $YahooFinance$ from 2009 and 2010. We collected daily S&P and NASDAQ indices. We also collect historical data on NASDAQ of seven large technology companies, including Apple, AMD, Cisco, Intel and Microsoft; and historical data on NYSE of seven leading banks, including , Citi Group, Morgan Stanley, JP Morgan, Barclays PLC, and Bank of America Corp. There were 503 trading days in 2009 and 2010. Stocks of all of these companies were fairly active during during 2009 and 2010.

Financial News Date The primary data source is New York Times corpus in XML format, which contains all news from 2009 to 2010. There are approximately 100,000 news articles for 500 trading days.

Tools We used Java to parse XML documents and remove stopping words. Topic model of this research is modified from a open source C++ implementation of LDA. The project is available at http://code.google.com/p/plda/.

4.2 Experiment Result

ADFIDD Using algorithm described in Chapter 3, we calculate the estimate indices of trading volume and market trend for 10 stocks mentioned earlier. And then we the correlation between estimate indices and actual trading volume and market trend.

Table 1. correlation between ADFIDF feature and market information in IT Industry

	AAPL	AMD	CSCO	INTC	MSFT
Volume	0.113	-0.18	0.154	0.016	0.059
Market trend	0.24	0.031	-0.123	-0.066	0.171

Table 2. correlation between ADFIDF feature and market information in Bank Industry

	BAC	BLC	C	MS	JPM
Volume	-0.097	0.069	0.054	0.216	0.146
Market trend	-0.182	-0.252	-0.043	0.106	-0.087

From Table1 and Table2, we can see that trading volumes of 8 out of 10 stocks appear to be positively related with estimated volume indices while market trends are not correlated with estimate indices at all. The result meets our expectation. Volume of the market is generally easier to predict. Whenever a stock is effected by a related news, the volatility will be increased because people might see potential profit or risk from the news information. Either interpolation will motivate people to trade on this stock, which increase the market volatility and trading volume. The direction of market, however, is not directly related to either volatility or volume. Therefore, market trend is harder to predict.

Although positive correlation exits between trading volume and estimate indices, it are not very strong. For all stocks, correlations are less than .2 and most of them are less than .15. The correlation between volume and features is much weaker than the one between features and volatility from previous research work. Rank documents' relevance to each stock and then use it to adjust estimate indices might improve the performance because in previous research, they used only news that are relevant to targeting stocks, while in our research, due to limited number of news for each stocks on a day, we used all documents of that day to select the feature.

In addition, market volume of bank industry appears to be slightly more positively correlated with estimate indices than IT industry. The result is reasonable. IT firms are

usually more stable because their work is not directly related to financial market. Banks, however, are more sensitive to market movements and human activity.

Topic Model After building the topic model, we calculate score for each word and topic as described in Section 3. There were more negative scored topics than positive scored topics for both 8 topics and 20 topics model especially when TF-IDF weighting is not used. This should reflect the real market since during the 1999 and 2000, the market has more days going down than going up. Topic models with 5 topic is a little sparse. So the prediction result when topic number equals 5 might not be vary reliable.Finally, we run topic model on test data and get the distribution of topics and then we calculate the binary score for both training data and test data.

Table 3. Experiment result: market trend of NASDAQ

	with TF-IDF weighting				without TF-IDF weighting			
	Training		Test		Training		Test	
# of topics	20	10	20	10	20	10	20	10
Success	53340	864	54535	865	53680	801	53304	818
Fail	55767	826	54572	825	55427	889	55803	872
%	0.511	0.489	0.500	0.512	0.492	0.474	0.489	0.484
correlation	0.059	-0.039	0.025	0.014	-0.008	0.043	-0.08	0.048

Table 4. Experiment result: market trend of Bank Industry

	with TF-IDF weighting				without TF-IDF weighting			
	Training		Test		Training		Test	
# of topics	20	10	20	10	20	10	20	10
Success	54528	850	52816	864	54635	808	52850	875
Fail	54579	840	56291	826	54472	882	56257	815
%	0.500	0.503	0.511	0.484	0.501	0.478	0.484	0.518
correlation	-0.030	-0.015	-0.014	0.040	-0.013	-0.036	-0.047	0.026

Table 5. Experiment result: market trend of IT Industry

	with TF-IDF weighting				without TF-IDF weighting			
	Training		Test		Training		Test	
# of topics	20	10	20	10	20	10	20	10
Success	52809	847	53999	870	53558	877	53243	881
Fail	56298	843	55108	820	55549	813	55864	809
%	0.501	0.484	0.515	0.495	0.519	0.491	0.488	0.521
correlation	0.047	0.020	-0.009	0.020	-0.036	0.038	0.014	-0.04

Table 6. Experiment result: volume of NASDAQ

	with TF-IDF weighting				without TF-IDF weighting			
	Training		Test		Training		Test	
# of topics	20	10	20	10	20	10	20	10
Success	53616	803	53597	843	53991	848	53120	831
Fail	55491	887	55510	847	55116	842	55987	859
%	0.491	0.475	0.491	0.499	0.495	0.487	0.502	0.492
correlation	0.08	-0.029	-0.015	-0.03	-0.031	-0.014	-0.038	0.027

Table 7. Experiment result: volume of Bank Industry

	with TF-IDF weighting				without TF-IDF weighting			
	Training		Test		Training		Test	
# of topics	20	10	20	10	20	10	20	10
Success	53647	842	53488	884	54583	814	53427	883
Fail	55460	848	55619	806	54524	876	55680	807
%	0.492	0.498	0.523	0.490	0.500	0.482	0.490	0.522
correlation	0.01	-0.014	-0.043	-0.03	-0.014	0.033	-0.104	-0.048

Table 8. Experiment result: volume of IT Industry

	with TF-IDF weighting				without TF-IDF weighting			
	Training		Test		Training		Test	
# of topics	20	10	20	10	20	10	20	10
Success	52814	808	52935	894	53676	835	54202	885
Fail	56293	882	56172	796	55431	855	54905	805
%	0.484	0.478	0.485	0.529	0.492	0.494	0.497	0.524
correlation	0.022	0.029	-0.037	-0.106	-0.017	0.005	-0.085	0.003

From the experiment result we can see that the prediction algorithm does not alway have a consistent performance. When using TF-IDF weighting algorithm to calculates the topic score, the experimental result of binary prediction is slightly better than not using TF-IDF. This means, TF-IDF weight is still a important factor when predicting market trends. In addition, when using 20 topics, the model performs a little better and is

more stable than using 10 topics. Constructing more topics when building a topic model is a good way to assign more detailed distribution of words to each topic, which could increase the accuracy of model. Performance of training data is almost the same as test data. This could be because of the large size of data, so that a it don't favor a even if it was traing data.

CHAPTER 5

CONCLUSIONS & FUTURE WORK

5.1 Conclusions

In conclusion, base on our experiment result, ADFIDD doesn't correlated with market trend but correlated with volume. This is probably because trading volume is more directly related to activeness of the market which is directly influenced by published news. Correlations maybe improved by ranking relevance of the news to each stock during the process. Banking industry is more positively correlated with ADFIDF features than IT industry.

In addition to studying the existing methods, we also proposed a new method where we use topic models to obtain topic-level representation of text and make prediction based on topics rather than keywords. While this new approach appears to be promising, our experiment results, however, failed to show that it can outperform other existing methods. Right now we are not sure whether topic model would improve performance of any prediction algorithm. The experiment results also show that TF-IDF weighting might be necessary in building the prediction model. Also, using more topics when trading model could improved the performance and stability of prediction. Business knowledge might be necessary when predicting market using pure text analysis algorithms. Grouping stocks together does not improve the performance of prediction

5.2 Further Work

Noise removal and news labeling There are massive news being published everyday, selecting which ones are more effective than others could improve the performance of model. For example, sports news shell weight less than finical related news. Using historical market data and news to rank the relevance of news or labeling breaking news might be a good way to select data. Labeling news with positive or negative tags manually and then using those tag to predict market trends and volatility is also interesting to look into. For example, similar as semantic classification of product reviews, if news can also be labeled with tags indicating how they will influent the market, then analyzing correlations become much simpler. However, the challenge is how could we get accurate news' reviews.

Better Representation of Market Data Because of some risk control policies, many trading firms will try to be "flat", no short sell and no holds of stocks, at the end of the day. So using closing price minus open price might not be the best representation of that day's market movement. Using average or median value of the closing hour could improve the prediction. Using data by the hour might also reduce the cause of trading behavior, however, news with accurate time label are needed which can be very challenging to get.

Business Knowledge Models of our research used only text mining techiques. Business interpretation of an article is also very important. Mentioned in Section 2. Experts of trading market construct a set of keywords which they think are important for moving markets. Words are grouped and assigned weights. Then they classify news articles base on their distribution over key words. This method can also be used to assist ADFIDD and topic models prediction method.

REFERENCES

[1] Timmons, H. (2006). A London Hedge Fund That Opts for Engineers, Not M.B.A.s. Retrieved on July 1, 2012, from http://www.nytimes.com/2006/08/18/business/worldbusiness/18man.html?_r=1&ex=1313553600&en=b2fee1b41c85af15&ei=5088&partner=rssnyt&emc=rss

[2] Haldane, A. G. (2010). Patience and Finance. Retrieved on June 15, 2012, from http://www.bankofengland.co.uk/publications/speeches/2010/speech445

[3] Hook, J., Lee, C. E., and Lauricella, T. (2012). Jobs Power Market Rebound. Retrieved July 1, 2012, from http://online.wsj.com/article/SB10001424052970203711104577200730710149216.html

[4] Gross-Klussmann, A. and Hautsch, N. (2011). When Machines Read The News: Using Automated Text Analytics to Quantify High Frequency News-implied Market Reactions. *Journal of Empirical Finance*, 18(2), 321-340

[5] Pan, Q., Cheng, H., Wu, D., Yu, J. and Ke, Y. (2010). Stock Risk Mining by News. Hong Kong: The Chinese University of Hong Kong Press

[6] Salton, G. and Zhang, Y. (1986). Enhancement of text representations using relateed document titles. *Information Processing & Management*, 22(5), 385-394

[7] Koppel, M. and Shtrimberg, I. (2004). Good News or Bad News? Let the Market Decide. *Computing Attitude and Affect in Text: Theory and Applications*, 297–301.

[8] Cheng, S. H. (2010). Forecasting the Change of Intraday Stock Price by Using Text Mining News of Stock. *2010 International Conference on Machine Learning and Cybernetics, ICMLC 2010*, 5, 2605-2609

[9] Shihavuddin, A.S.M., Ambia, M.N., Ardin, M.M.N., Hossain, Md.M., Anwar, A. (2010). Prediction of Stock Price Analyzing The Online Financial News Using Naive Bayes Classifier and Local Economic Trends. *ICACTE 2010-2010 3rd International Conference on Advanced Computer Theory and Engineering*, 4, 422-426

[10] Mittermayer, M. (2004) Forecasting Intraday Stock Price Trends with Text Mining Techniques. *Proceedings of the Hawai'i International Conference on System Sciences*, 3, 30064.2

[11] Schumaker, P. R. and Chen, H. (2009). Textual Analysis of Stock Market Prediction Using Financial News. *ACM Transactions on Information Systems*, 27(2), a12

[12] Sven, G. and Jan, M. (2011). An intraday market risk management approach based on textual analysis. *Decision Support System*, 50(4), 680-691.

[13] Lerman, K., Gilder, A., Dredze, M., Pereira, F. (2008). Reading the Markets: Forecasting Public Opinion of Political Candidates by News Analysis. *Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference 1*, 473-480

[14] Blei, D. M., NG, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022

[15] Hafez, P. (2009). How Does the Market React to News? Retrieved on July 1, 2012, from `http://www.sentimentnews.com/2009/06/how-does-market-react-to-news.html`

[16] Anderson-Weir, Charles H. (2010). How Does the Stock Market React to Corporate Environmental News? *Undergraduate Economic Review*, 6(1), Article 9

[17] Ruiz, E., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. *WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining*, 513-522

[18] Harvey, D. K. (2009). Forecasting the belief of the population: Prediction Markets, Social Media & Swine Flu. Retrieved June 30, from `http://www.mendeley.com/profiles/dan-harvey/`

[19] Kogan, S., Levin. D., Routledge, B. R., Sagi, J. S. and Smith, N. A. (2009) Predicting Risk from Financial Reports with Regression. *NAACL '09 Proceedings of Human Language Technologies*, 272-280

[20] Lowd, D. and Domingos, P. (2005). Naive Bayes Models for Probability Estimation. *ICML '05 Proceedings of the 22nd international conference on Machine learning*, 529-536

[21] Fuerbringer, J. and Norris, F. (2001). A DAY OF TERROR: THE MARKETS; Stocks Tumble Abroad; Exchanges in New York Never Opened for the Day Retrieved on July 1, 2012, from `http://www.nytimes.com/2001/09/12/business/day-terror-markets-stocks-tumble-abroad-exchanges-new-york`